基于文本挖掘和机器学习的股指预测与决策研究

戴德宝[1]，兰玉森[1]，范体军[2]，赵　敏[3]

（1. 上海大学　管理学院，上海　200444；
2. 华东理工大学　商学院，上海　200237；
3. 上海大学悉尼工商学院，上海　201800）

摘　要： 依据行为金融学理论，资本市场投资者的心理和行为对股票指数变动有重要影响。为此本文假设投资者情绪与股票指数存在一定内在作用机制，能预测股票市场整体价格变化。通过文本挖掘技术和情感分析方法生成积极和消极各三阶共六类投资者情绪时间序列数据；采用单位根检验、Granger 因果关系检验和因子分析等方法构建上证投资者情绪综合指数，并分别使用支持向量机和神经网络预测股票市场价格变化，进行假设验证。结果表明：利用网络股市论坛文本数据和股票交易数据构建的上证投资者情绪综合指数能够提高股指走势预测的精度，有利于政府、在线平台、上市公司和投资主体更好决策。

关键词： 投资者情绪；股票预测；文本挖掘；机器学习

中图分类号： F832.51　　**文献标识码：** A　　**文章编号：** 1005－0566(2019)04－0166－10

Stock Forecast with Investors Sentiment by Text Mining and Machine Learning

DAI De-bao[1], LAN Yu-sen[1], FAN Ti-jun[2], ZHAO Min[3]

(1. *Management School, SHU, Shanghai* 200444, *China*;
2. *Business School, ECUST, Shanghai* 200237, *China*;
3. *SILC Business School, SHU, Shanghai* 201800, *China*)

Abstract: According to the theory of behavioral finance, investors' psychological and behavior have important influence on the trend of stock market index. For this reason, this paper assumes that investors' sentiment is inherently associated with stock market index, which can predict the overall price change of a stock market. In this research, six kinds of the time series of investors sentiment are constructed by means of text mining technology and emotion analysis methods while Shanghai stock exchange composite investor sentiment index (*SSECISI*) is created by using unit root test, Granger causality test and factor analysis. The SVM and the neural network model are used to predict the stock market index to verify the correctness of the hypothesis. The results show that the *SSECISI* constructed by the forum mood and stock transaction data can improve the forecast precision of stock market index and enable government, online platforms, listed companies and investors to make decision better.

Key words: investors sentiment; stock forecast; text mining; machine learning

收稿日期： 2018－10－27　　**修回日期：** 2019－04－02

基金项目： 国家自然科学基金重点项目(71431004)；教育部人文社会科学研究规划基金项目(17YJA880014)。

作者简介： 戴德宝(1972－)，男，河南固始人，上海大学管理学院副教授，博士，研究方向：金融信息，决策支持。

一、引言

《世界互联网发展报告 2018》和《中国互联网发展报告 2018》蓝皮书数据显示: 2017 年,中国数字经济总量达 27.2 万亿元,对 GDP 增长贡献率达 55%,全球数字经济规模达 12.9 万亿美元,中国位居全球第二。以互联网为代表的信息技术和人类生产生活深度融合,引领创新,驱动转型。社交平台作为数字经济呈现形式之一,现已是消费者或投资者交换观点、情感和知识的重要渠道。与调查问卷、档案数据和访谈记录等信息源相比,社交平台数据能够规避传统信息收集方式的滞后、缺失和高投入等弊端,具有用户基数大、社交性强、涉入性高、响应速度快等优势。借助博客、微博和论坛等不同社交平台在线文本,利用文本挖掘和情感分析技术可以研究许多相关主题[1]: 使用在线评论分析结果减少网络购物不确定性和风险[2],使用社交平台用户的产品感知和意见挖掘结果优化产品品质和提高品牌价值[3],发现学习社区对学习效果的影响与促进作用[4],检验在线投资者情绪与资本市场的关联状况[5-6]。网络社交平台已成为在线商品和服务交易数据观察利用空间。党和政府给予高度评价、期望和要求,十九大报告提出"贯彻新发展理念,建设现代化经济体系"。"互联网 + 金融"促进金融体制改革,允许优质企业申办网络银行[7-8],开放小额贷款平台,允许互联网企业施行消费贷款,利用用户原创内容(user generated content,UGC)分析用户行为和预测市场趋势。

金融市场规律研究或趋势分析有助于金融机构和投资者防范金融风险、增强现代金融监管并促进金融体系良性运转。股市分析技术证明资本市场有后验规律但难以把握未来,股价是否能够预测莫衷一是。由于新信息随机性和不可预知性,股票价格处于无规则行走模式,未来价格根本赌注是现在价格,预测准确率将不超过 50%。然而许多研究结果表明股价不遵循随机漫步理论,而是受公司财务情况、宏观经济指标和历史交易数据等众多因素影响,可以使用多维度的数据预测[9],股票走势预测准确率到达 56% 即为满意[10]。金融学、心理学和行为学等结合派生的行为金融学[11]认为股票价格并非只由企业内在价值决定,很大程度上受投资者心理和行为影响。基于投资者情绪的股价预测研究框架主要涵盖以下三个方面。

(1) 情绪资源。一是网上新闻: 金融新闻否定句与股价相关关系[12]以及纽约时报和 40 个世界金融指数联系研究有力支持行为金融学新经济范式作用[13]。二是社交媒体资源: 社交平台的投资者文本情绪影响股价[14]。由于微博推文内容无法聚焦和用户地理位置无法确定等缺陷[15],近期研究选择股民聚集度高、话题专业性强、情绪传递性快的财经论坛(如 StockTwits[16]、Yahoo 财经网[10,17]和东方财富网[18-20])挖掘投资者情绪。

(2) 情绪指标。一是与数量相关指标: Google 搜索量(Search Volume Index,SVI)的增长能够预示未来两周股价上涨[21],股吧社区发帖量影响股价[20]; 二是与情绪相关指标: 各类社交媒体整体情绪与股票回报和投资风险有关,且优越于传统媒体[22]。影响股价的情绪可分为六个维度: Calm、Alert、Sure, Vital、Kind 和 Happy 等[6],或者五个维度: 强烈买入、买入、中性、卖出和强烈卖出等[18]。

(3) 预测对象。一是个股股价走势: 多家公司 Twitter 情绪和异常股票回报相关[5],投资者浏览行为及情绪变化能够有效预测股票[23],投资者情绪通过网络自媒体传播会影响多只股票收益[20]; 二是股票价格指数(即股指)预测: 沪深 300 指数探究投资者情绪与股价存在因果关系[15,19]。常见预测股指包括道琼斯指数(DJIA)[6]、标准普尔指数(S&P500)[14]、上证指数(SSEC)[24]等。

金融市场预测方法包含经典统计学的多元回归模型[18,22]、自向量回归模型[25]以及支持向量机(support vector machine, SVM)[10,14,24]、神经网络[6,14,25]、随机森林[14]等现代机器学习方法,SVM 和 BP 神经网络应用最多。传统回归分析以严格假设和充足先验为前提,难以构建有效金融预测模型,机器学习能够自主学习反复改善和优化算法,结果满意[26]。其他如 Adaboost、LinearSVC 等方法逊于 SVM 和 BP 神经网络对复杂非线性问题

的处理。

许多基于文本挖掘的金融市场关联或预测文献研究直接将单一维度情绪变量(积极情绪或消极情绪)直接加入模型,而且少有对非线性和高噪音情绪数据进行处理,容易验证是否与金融市场关联,难以取得较好的预测效果。本文通过抓取东方财富股票论坛数据,借鉴天气或事件的金融关联分析过程[27],不仅剔除中性或噪音数据,而且选取相关性强的情绪数据参与投资者情绪指数设计,基于情绪数据和股指数据非线性特征,利用 SVM 和 BP 神经网络两类模型进行股指预测,证明投资者情绪与股指存在内在联系,并且预测高效,以期为投资者、上市公司和政府监管部门的决策支持提供良好参考价值。

二、基于文本挖掘和机器学习的股指预测

基于文本挖掘和机器学习的股指预测内容包括股指和情绪两种数据的预处理和平稳性检验、预测组合指数构建及数据生成、常用两种股指预测的机器学习算法检验等四个部分。

(一)股指数据获取与情绪数据预处理

(1)情绪数据获取与预处理。投资者情绪文本数据源于东方财富网股吧论坛实战吧,使用 Python 共抓取帖子 368586 条,跨度:2016 年 7 月 19 日至 2017 年 12 月 29 日。通过编写帖子清洗规则剔除不能表达投资者情绪的主题帖,共保留帖子 217445 条。清洗规则包括图片(无文字)、链接(无文字)、乱符(无意思)和实盘组合(系统自动生成)等四种相关类型;文本情绪分类方面,利用基于词典的中文情感分析方法[28]对帖子情感打分。词典由情感词、程度副词和否定词三类词汇组成,根据式(1)计算帖子综合情感得分。情感词包括通用情感词典和专用情感词(阴跌、利好、诱多、狗庄和割肉等)。

$$PostScore = Wr \cdot Wm\left\{\sum_{i=1}^{m}\left[\left(\prod_{j=1}^{n} Wd_j\right)\cdot\left(\prod_{jj=1}^{nn} Wn_{jj}\right)\cdot Ws_i\right]\right\} \quad (1)$$

其中,$PostScore$ 为情感综合得分,m 为一个帖子标题的情感词数目,n 和 nn 分别为第 i 个情感词前面程度副词数量和否定副词数量;Ws、Wm 和 Wr 分别为对应帖子标题的各情感词分值、各标点符号分值和反问词分值;Wd 和 Wn 分别为对应情感词前面的程度副词分值和否定副词分值。

本实验主要研究积极与消极情绪参与的股指预测,将不同情绪帖子数量按天归类处理,得到一般积极、中度积极、高度积极、一般消极、中度消极、高度消极六个具有情绪倾向的时间序列数据[24],分别计入变量 PI、PII、$PIII$、NI、NII、$NIII$。

(2)股市交易数据获取。上证指数(000001)交易数据导出自通达信金融终端,时段自 2016 年 7 月 19 日至 2017 年 12 月 29 日 356 个交易日的历史信息:收盘价(*CLOSE*)、开盘价(*OPEN*)、最高价(*HIGH*)、最低价(*LOW*)、成交量(*VOL*)和成交额(*AMO*),综合考虑相关系数矩阵结果及变量实际意义,选取收盘价表示上证指数数据(*SSEC*)。

(二)股指数据与情绪数据平稳性检验

(1)数据标准化。为消除股票交易数据和投资者论坛情绪数据间的量纲关系,提高数据可比性,需对两类数据按照式(2)进行标准化(Z-Score)处理,μ 为样本数据均值,σ 为样本数据标准差。

$$z = (x - \mu)/\sigma \quad (2)$$

(2)单位根检验。是通过对时间序列矩的随机游走检验排除统计数据的偏误及模型的伪回归,保证预测模型的稳定性,不存在单位根则时间序列平稳。本文选用 ERS(Eiliot, Rothenberg and Sock Point Optimal Test)检验单位根,避免检验包含常数项和趋势变量项。

检验结果(见表 1)表明:*SSEC*、*OPEN*、*HIGH*、*LOW* 四个时间序列变量的 ERS 检验统计值大于在 10% 置信度下的临界值,这些时间序列变量包含单位根,是非平稳的。

表 1 时间序列的单位根检验

指标	*SSEC*	*OPEN*	*HIGH*	*LOW*	*VOL*	*AMO*	*PI*	*PII*	*PIII*	*NI*	*NII*	*NIII*
ERS 值	17.2325	13.5814	21.4395	15.1403	0.8254	0.5192	4.2104	2.4557	0.5254	1.8937	1.0911	0.2846

注:表 1 和表 2 中,当显著性水平为 1%、5% 和 10% 时,检验临界值分别为 1.972、3.240 和 4.447。

(3) 差分时间序列单位根检验。将所有变量按照式(3)进行一阶差分运算后得到新的序列变量，分别记作：*DSSEC*、*DOPEN*、*DHIGH*、*DLOW*、*DVOL*、*DAMO*、*DPI*、*DPII*、*DPIII*、*DNI*、*DNII*、*DNIII*，X_t和X_{t-1}分别为 t 和 t－1 时段变量值。

$$D(X) = X_t - X_{t-1} \quad (3)$$

对一阶差分后各时间序列进行单位根检验(见表 2)发现：ERS 统计值均小于在 1% 置信度下的临界值，最大 ERS 值为 0.233，各时间序列趋于平稳状态。

(三) 选取相关数据生成组合指数数据

(1) 相关性分析。上证指数历史交易数据变量差分后采用 Pearson 相关分析法发现各变量相互影响且存在相关性(见表 3)，可进行有效的股指预测。本文将选取 *DOPEN*、*DHIGH*、*DLOW*、*DVOL*、*DAMO* 五个变量构造上证交易组合指数。

表 2　差分时间序列的单位根检验

指标	*DSSEC*	*DOPEN*	*DHIGH*	*DLOW*	*DVOL*	*DAMO*	*DPI*	*DPII*	*DPIII*	*DNI*	*DNII*	*DNIII*
ERS 值	0.1836	0.1836	0.1401	0.1734	0.1257	0.1281	0.1670	0.0372	0.0437	0.0037	0.2326	0.0091

表 3　各变量间的相关系数矩阵

差分交易指标	*DSSEC*	*DOPEN*	*DHIGH*	*DLOW*	*DVOL*	*DAMO*
DSSEC	1					
DOPEN	0.085	1				
DHIGH	0.557**	0.704**	1			
DLOW	0.599**	0.557**	0.626**	1		
DVOL	0.064	0.314**	0.476**	−0.052	1	
DAMO	0.087	0.356**	0.516**	0.005	0.979**	1

注：** 表示在 1% 水平(双侧)上显著相关。

(2) Granger 因果关系检验。假设投资者易受其他投资者情绪影响而选择非理性投资，需要对上证指数和六组投资者情绪时间序列进行 Granger 因果关系检验，分析和验证投资者情绪变化是否关乎市场波动，是否能够预测股指信息[28]。Granger 因果关系检验解释是：变量 x 是否为变量 y 的产生原因可以观察当前 y 在多大程度上能被过去 x 解释。如果 x 滞后值能提高 y 解释程度，说明 x 有助于 y 的预测，y 是由 x 的 Granger 因果引起[29]。尽管 Granger 因果关系检验结果不等于实际因果关系，但本文目的不是测试实际因果关系，而是测试投资者情绪时间序列是否存在上证指数时间序列的预测信息。

除去双休日和法定节假日，股票实际交易日为一周 5 天，滞后期可分别选取为 1 天到 5 天。Granger 因果关系检验结果(见表 4)表明：一般积极情绪(*DPI*)在滞后 1 天到滞后 3 天与上证指数存在较为显著的 Granger 因果关系(p 值 <0.04)。图 1 为 $DPI(t-3)$ 和 $DSSEC(t)$ 两个时间序列对比图，阴影部分表示 *DSSEC* 与滞后 3 天的 *DPI* 时间序列存在重叠或者有相同趋势。无论是 Granger 因果关系检验结果还是时间序列图，都可从中得出一般积极情绪与上证指数存在显著相关关系，即 *DPI* 可用于预测上证指数。

表 4　Granger 因果关系检验结果

滞后天数	*DPI*	*DPII*	*DPIII*	*DNI*	*DNII*	*DNIII*
1	0.0313*	0.5365	0.5076	0.1160	0.1114	0.9777
2	0.0392*	0.2329	0.7138	0.0860	0.0788	0.9891
3	0.0215*	0.0598	0.4623	0.0559	0.1404	0.4959
4	0.0424*	0.0943	0.5424	0.0814	0.1033	0.5571
5	0.0470*	0.0694	0.6458	0.1316	0.1576	0.6503

注：表格中的数值为 p 值，表示“检验行名称不是 *SSEC* 因果关系”，其中* 表示在显著性水平为 5% 下显著。

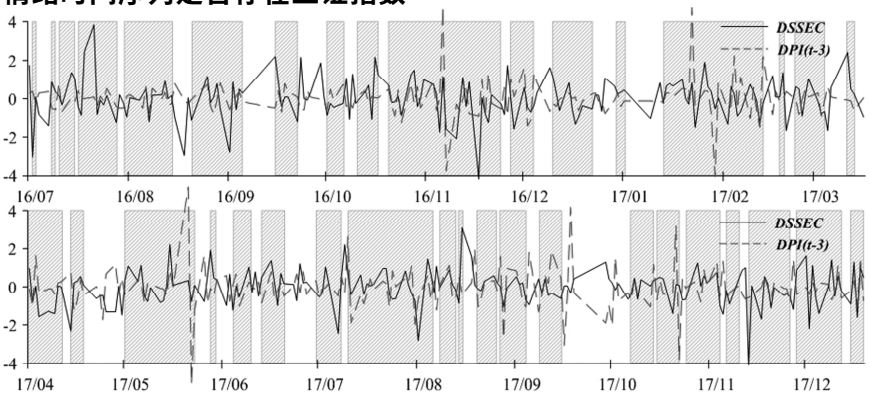

图 1　上证指数与一般积极时间序列情绪对比图

注：灰色背景部分为上证指数和滞后 3 天的一般积极情绪走势相同区域。

(3) 因子分析和指数构建。本文选用多维度指标方法避免投资者情绪使用单一指标代理变量的代理有偏和信息不足问题,通过对六个变量(*DOPEN*、*DHIGH*、*DLOW*、*DVOL*、*DAMO*、*DPI*) 因子分析得出上证投资者情绪综合指数(*SSEC* Investor Sentiment Index,*SSECISI*)。为验证投资者情绪对股指预测的高效性,从 *SSECISI* 中剔除 *DPI*,仅利用 *DOPEN*、*DHIGH*、*DLOW*、*DVOL*、*DAMO* 五个变量构建上证交易组合指数(*SSEC* Portfolio Index,*SSECPI*)。使用主成分分析法先对因子载荷矩阵进行方差最大正交变换求得因子得分(式 4) 和方差贡献率(见表 5),然后根据因子得分和方差贡献率的加权平均(式 5) 获得 *SSECPI* 和 *SSECISI* 数据[29]。

$$F_j = \beta_{j1}X_1 + \beta_{j2}X_2 + \cdots + \beta_{jp}X_p, j = 1,2,\cdots,m \quad (4)$$

其中,F_j 为因子 j 的因子得分,β_{jp} 为成份 X_p 的因子得分系数。

$$F = (F_1V_1 + F_2V_2 + \cdots + F_jV_j) / \sum_{j=1}^{m} F_j, j = 1,2,\cdots,m \quad (5)$$

其中,F 为综合得分,即本文构造的指数,V_j 为因子 j 的贡献率。

(四) 两种预测方法股指预测检验

(1) 建模预测。Granger 因果关系检验表明情绪数据含有股指走势信息,可以选取预测方法构建预测模型。基于情绪数据和股指数据的非线性特征以及机器学习模型的良好非线性数据处理能力[26],本文选取构建机器学习预测模型,希望能很好解释投资者情绪与股指趋势的非线性关系。因为 SVM 和 BP 神经网络都能处理非线性数据而又各有所长,本文则采用 SVM 和 BP 两种方法对比验证,避免随机和偶然,以发现更好的适用方法。

建模前先将 356 个交易日的上证指数和投资者情绪数据作为样本,根据不同时长分为三组:第 1 组时长 18 个月,起始日期为 2016/07/19,样本量、训练集、测试集分别为 365、267、89 天;第 2 组时长 9 个月,起始日期为 2017/04/05,样本量、训练集、测试集分别为 185、136、49 天;第 3 组时长 4.5 个月,起始日期为 2017/08/16,样本量、训练集、测试集分别为 93、72、21 天。为检验上证投资者情绪综合指数对上证指数收盘价预测结果的影响,特设计 3 组不同排列的输入变量:P_0,P_{SSECPI} 和 $P_{SSECISI}$。P_0 选取上证指数交易日 t 前 3 天的收盘价($SSEC_{t-3,2,1}$),P_{SSECPI} 和 $P_{SSECISI}$ 是在 P_0 基础上分别加入滞后 1 天至 3 天的交易组合指数($SSECPI_{t-3,2,1}$) 和投资者情绪综合指数($SSECISI_{t-3,2,1}$),如式(6) 所示:

$$\begin{aligned} P_0 &= \{SSEC_{t-3,2,1}\} \\ P_{SSECPI} &= \{SSEC_{t-3,2,1}, SSECPI_{t-3,2,1}\} \\ P_{SSECISI} &= \{SSEC_{t-3,2,1}, SSECISI_{t-3,2,1}\} \end{aligned} \quad (6)$$

本实验使用 BP 神经网络和 SVM 两种方法对三组输入向量分别实验。实验前通过归一化处理消除变量量纲,将数据归于 [0,1] 之间,如式(7) 所示:

$$X^* = (X - X_{\min})/(X_{\max} - X_{\min}) \quad (7)$$

其中,X_{max} 和 X_{min} 分别为测试集中各变量的最大和最小值。

运行环境与参数设置方面,BP 神经网络:Kosmogorov 定理证明合理结构和恰当权值的三层前馈网络具备逼近任意连续函数能力,故隐含层层数皆设置为 1;根据反复实验和择优原则,设置隐含层神经元个数为 6;学习速率为 0.01,最小训练误差目标为 0.001,最大迭代次数为 100。SVM:数值型变量分类方式采用 ∈ 类支持向量回归机(EPS-SVR),Kernel 非线性映射函数(核函数) 选取双曲正切函数(Tanhdot),核参数为 1/k(k 为特征向量的个数),惩罚参数 C 为 1。

表 5 *SSECPI* 与 *SSECISI* 因子分析结果

综合指数	主因子	*DOPEN*	*DHIGH*	*DLOW*	*DVOL*	*DAMO*UNT	*DPI*	特征值	V_j方差(%)
SSECPI	FPI_1	0.266	0.425	-0.180	*0.985*	*0.976*	-	2.207	44.145
	FPI_2	*0.825*	*0.824*	*0.902*	0.095	0.152	-	2.205	44.108
SSECISI	$FISI_1$	*0.840*	*0.875*	*0.868*	0.229	0.282	-0.114	2.371	39.521
	$FISI_2$	0.180	0.309	-0.286	*0.947*	*0.935*	*0.349*	2.101	35.020

采用走势准确率(*Direction*,向上或向下)对SVM和BP神经网络的预测精度进行评价。其定义如下:

$$Direction = \sum_{t=2}^{n} \hat{y}'_t / \sum_{t=2}^{n} y'_t, t = 2,3,\cdots,n \tag{8}$$

其中 $y'_t = \begin{cases} 0, y_{t-1} \neq y_t \\ 1, y_{t-1} = y_t \end{cases}$; $\hat{y}'_t = \begin{cases} 0, \hat{y}_{t-1} \neq \hat{y}_t \\ 1, \hat{y}_{t-1} = \hat{y}_t \end{cases}$,

$t = 2,3,\cdots,n$; y_t 和 $\hat{y}_t$ 分别为为第 t 时刻实际值和预测值,n 为预测的时段数。

(2)算法准确率检验。本文分别采用BP神经网络和SVM方法对 P_0、P_{SSECPI}、$P_{SSECISI}$ 三组样本进行实验得股指走势准确率对比结果(见表6):SVM预测准确率普遍优于BP神经网络;$SVM-P_{SSECISI}$ 模型预测准确率在59%-70%,大于股指预测准确率满意值56%[12],具有有效性;两种预测方法的平均预测准确率发现 $P_{SSECISI} > P_{SSECPI} > P_0$,说明上证交易组合指数模型比纯股指预测模型的预测准确率高,而上证投资者情绪综合指数模型又比上证交易组合指数模型的预测准确率更高。综合结果表明使用机器学习进行股指预测,SVM方法下的投资者情绪数据参与的综合预测模型最优。

表6　BP神经网络与SVM模型走势准确率(%)

预测模型	第1组	第2组	第3组	平均值
$SVM-P_0$	51.69	54.17	45.00	50.38
$SVM-P_{SSECPI}$	58.43	56.25	65.00	59.89
$SVM-P_{SSECISI}$	59.55	62.50	70.00	64.02
$BP-P_0$	49.44	45.83	45.00	46.76
$BP-P_{SSECPI}$	51.68	45.83	55.00	50.84
$BP-P_{SSECISI}$	53.93	47.92	60.00	53.95

(五)预测效果与技术分析

(1)预测效果分析。本实验中SVM预测效果优于BP神经网络,可能原因是BP神经网络易陷入局部最优的欠拟合和过拟合问题,而SVM核函数能将复杂非线性问题转变为线性问题,增强鲁棒性;$P_{SSECPI} > P_0$ 的原因在于市场交易的收盘价不由单一历史收盘价决定,而是历史多期多指标(开盘价、最高价、最低价、成交量、成交额)的共同作用,类似于量价技术分析模型(Trade Amount Per Index,TAPI)效果;$P_{SSECISI} > P_{SSECPI}$ 的原因是多指标数据综合效应依然不能完全准确决定市场趋势,不能全面反映投资者的主客观决策依据。资本市场投资决策的复杂性说明需要补充更多的信息来源(如投资者情绪数据)才能尽量准确预判市场趋势;时长对比结果并不全是第3组>第2组>第1组,但第3组最优,第2组在所有BP神经网络算法下低于第1组,在SVM算法下 P_{SSECPI} 效果低于第1组,说明时长在预测中的重要性,第3组单季度范围数据预测效果可能因为无周期成分扰动而好于另外两组,年度数据与三个季度数据则出现预测准确率排序不确定现象;另外,预测准确度还与数据采集和预处理相关,清洗规则、标准化方法和情感词典完备性都会影响在线情绪数据质量。

(2)文本挖掘技术。以文本格式为主导的网络非结构化数据据称占据全球全部数据量80%以上,包括电子邮件、文件、报告、表格、通话记录、新闻稿、博客、微博、微信、问答、论坛、评论等,而纯数字化数据占比较少。文本挖掘成为新型商业分析需求技术,用以观察各类商业行为及其效果。本文预测效果分析先决条件就是文本挖掘系列技术:文本数据采集和清洗、文本数据分词、文本情感词典构建、文本数据情感打分、情感数据标准化等。如舆情和评价等其他文本数据一样,一方面,金融论坛情绪数据获取与加工过程虽然没有太大的技术难度,但会遇到前所未有的相应领域数据处理规则问题:数据采集规则、数据清洗规则、情感词判分规则、情感语句判分规则等。这些已有的文本数据加工规则都称不上完善或标准,目前还需要根据具体场景生成相关参数。另一方面,预测只能利用部分数据成分,而且是参与预测。获取文本时序数据后,再进行标准化后就可以参与分析和预测。本文在预测前还做对数据进行平稳性检验和相关分析,发现股指数据和加工所得的情绪数据都存在较大波动(非平稳性),转而思考使用差分数据,检验合规后进行相关分析,结果是一般积极情绪与股指有明显的相关。然后使用相关文本数据成分与股指其他指标组合构建新预测指数数据,而不是直接使用文本数据预测股指趋势。

(3) 机器学习预测技术。机器学习技术用于解决常规非线性问题,本文股指与文本两样数据都是非线性数据,不宜使用平滑类预测模型,而是选取 BP 神经网路和 SVM 两种常用机器学习模型进行股指预测,并发现更为适用的模型,结果是 SVM 算法优于 BP 神经网络,其他应用场景也可能相反。为观察时长影响,在预测过程分别使用三组时长不等数据对比试验,结果是短时预测效果更好。这说明基于文本非线性数据的预测研究需要考察方法、模型和时长等多维情形,更为复杂的数据可以采用机器学习与小波分析相结合预测。针对复杂的非线性数据源,预测的科学化、严谨性还需要更好的基准数据库和算法才能实现。科大讯飞人为参与机器同传事件说明机器学习目前还不具备理想的算法,要求机器实时随机同传翻译则忽视个性化语音和专业化词汇训练过程。如无大量语料库作用,机器学习难以胜任无规律的随机问题(未加训练的方言、术语和外来词等)。如果允许预先降噪和优化原始数据,滞后机器学习就会更好。另外,机器学习今天被广泛地应用于人工智能,实现途径就是完善地专业数据库和场景适用算法,诸如可接受的网络翻译和语音识别等普适性业务以及多数据源的投资理财服务等。

三、结论

通过抓取网络论坛情绪文本,提取金融专业词汇进行文本挖掘,实现文本挖掘数据的专业化和精准化;应用关联分析方法构建投资者情绪综合指数,消除直接使用情绪数据进行预测的有偏性;利用机器学习方法设计良好的股指预测模型,提升股指走势预测准确性,证明基于 SVM 的上证投资者情绪综合指数模型进行股指预测更加有效。

在线情绪数据可不可用?怎么利用?怎样用得更好?"可不可用"其实还是认识问题:主观性、随意性和主体差异性综合形成在线情绪数据的复杂性,情感词汇量化精准性影响在线情绪数据测度的科学性。在线情绪数据为现代研究接受与采用的主要原因是规模上超越局部复杂性和科学性的大数据宏观统计规律。"怎么利用"问题是要超越传统科学的因果律以大数据思维发现事物内在或外在关联性。在线情绪数据已被研究者用于数据挖掘,发现和验证市场规律,预测市场走势。专业数据公司和数据拥有者已开始使用在线大数据对用户开展跟踪画像、精准推荐、辅助产品和服务设计、市场定价等诸多行为决策;在线情绪数据要"用得更好"前提是:建构包容网络语言的数据化、科学化和动态化专业词库,使用结构化界面设计记录网络用户结构化数据(星级、关键词、摘要、数据图片),通过文本分析算法自动生成关键词,应对现阶段人工智能技术还未完全成熟的情况。

四、面向不同主体的决策支持建议

数据分析和决策支持离不开国内外经济形势研判,中美贸易摩擦逐渐深入和激烈,科技和金融是中美最大差距领域,也是增强我国经济驱动力的两个方向:硬策略和软策略。互联网技术学习与应用最为成功,主要归因于我国政府对此因势而谋、应势而动和顺势而为的默许、鼓励、支持和管控。金融市场虽与市场经济同时开启,但未在经济总量大幅攀升中获取经验,历经多次股灾,投资者、上市公司和监管部门依然存在非理性行为。十九大报告强调我国当前三大攻坚战:防范和化解重大风险、精准脱贫、污染防治,以解决经济快速发展引致的潜在和显性的宏观大问题。后两项解决三农和环境问题,消除贫困和增加消费,改善环境和提高生活品质。重中之重的是重大风险问题,包括金融失控风险、结构失衡风险、生产过剩风险以及多种风险组合形成的整体系统风险。因此,资本市场各个主体和服务支持者(在线平台和专业数据企业)要充分利用各类大数据,顺应国家和社会需求,积极稳定地投资该投资的,支持该支持的,管制该管制的。积极收集网络用户的声音和挖掘网络用户需求,汲取经验,预判未来,理性决策,防范各类金融风险。

第一,分析和利用在线投资者情绪数据,防范社会金融系统风险,保障市场健康发展。历次金

融危机说明资本市场有其自身的周期律，经济过热、流动失控、技术瓶颈、国家竞争和资本操控等复杂成因的单一或综合作用会导致一国或多国金融系统风险，监管部门需要将在线投资者情绪数据和行为金融学研究成果纳入市场监管新依据。仅采用交易数据甄别扰乱市场的违法违规操作不具备普遍监管效果。监管部门分析和利用在线投资者情绪数据：了解广大投资者对于资本市场整体态度和舆情态势以及对于监管措施的意见和建议，追踪金融事件和极端问题，及时调整监管方向并快速切入监管相关市场主体。监管部门还能够从数据分析在线投资者情绪，监管和防范股市剧烈波动，杜绝个人或机构发表批量舆论操控股价。

第二，完善面向大数据的技术能力，防范平台技术安全风险，增加平台数据收益。如电商平台一样，社交平台正常运营需要设备和技术保证。阿里、百度、京东、腾讯、当当等国内著名互联网公司都发生过宕机事件，折射出因用户量、数据量和峰值要求的技术安全问题。目前用户消费、沟通交流和娱乐等生活习惯都已经网络化，势必增加了平台数据流量，同时也挑战平台承载能力。在软硬件技术保障的前提下，除了收割广告和流量收益外，平台企业利用在线投资者情绪数据还可以：在法律允许范围内售卖用户行为数据获利，采用外包或自行分析方式获得数据分析中间成果或最终成果并进行售卖获利。平台企业进行数据分析的优越性在于数据的完备性，有利于个股、单个投资者、板块和整体股指的深入和精准分析，趋势预测和荐股结果会更加让人信服。平台竞争本质上就是技术、服务和用户的竞争，继而是服务器群、数据量和数据分析与挖掘的竞争，保障安全，攫取数据收益。

第三，成立大数据分析部门，助力上市公司研判市场趋势，精准投融资决策。资本市场行情影响上市公司财务战略决策，利好行情会有更多资本进入，方便增发股票和加大融资，也方便购买股票和加大投资。在线情绪数据能够帮助上市公司判断投资者对资本市场行情的主观评价与投资愿望，相关研究结果有助于上市公司判断资本市场行情，及时做好融资和投资决策。新建大型或小型社交网站或在著名社交网站平台开设企业专栏用于发现投资者的情绪信息和评价细节，并做好公司运营层面的管理与控制，通过积极的经营战略和积极的在线承诺防止相关负面情绪扩大化，保持良好声誉和品牌价值；有条件的上市公司建议成立大数据部门，招聘数据分析与挖掘人才，实现多源数据分析和利用的专业化和科学化，形成更为精准的投融资决策；小型上市公司可通过多种渠道购买在线情绪数据或者数据分析结果，观察市场，了解自己，把握先机。

第四，关注在线情绪数据和相关成果，增强个体投资合理性和稳健性，避免盲从风险。投资成为人们日常生活关键诉求，然而普遍存在一种“赌徒式”投机心理和“传销式”操作模式，无视交易数据、基本面数据和资本市场规律，缺乏对在线情绪数据的观察、分析与思考。非法股评专家、荐股师和金融衍生品的推销者利用微信群或 QQ 群诱导盲目的投资者。大量股民的非理性为个人或机构提供操纵股票的信心而导致股市剧烈波动，形成监管难度和散户损失。因此，个体投资者需要关注网络上其他投资者情绪数据和相关研究成果，辅助其他投资技术方法，参照基本面数据和交易走势数据，利用在线情绪数据的共识性投资态度和倾向，进行合理投资决策，避免投资过热和消极投资。

第五，理性对待人工智能热，优化资本布局，遵循技术与商业协同发展规律。资本布局首要追求是高回报，也易在经济热度上迷信“高风险”。普华永道预测：2030 年，中国 GDP 将达 38 万亿美金，有 7 万亿美金为人工智能（AI）驱动。高盛预测：2025 年，全球 AI 金融服务规模达 340－430 亿美元，AI 零售业规模将会高达 540 亿美元。国际权威机构 CB Insights 统计：2017 年，全球范围内有 152 亿美元投资进入 AI 领域，中国公司为 73 亿美元，占比 48%，位列第一。2017 年被称为 AI 商业

化元年。与之相反的数据是腾讯研究院的 AI 研究报告:中美倒闭 AI 企业总数已超过 50 家,AI 企业将迎来“倒闭潮”。原因是一些急功近利的资本误入商业上的“伪创新”和“伪概念”,无视或不清楚“自动”、“智能”与“智慧”的区别。投资主体和支持平台都需要理性认识技术演进和拓展规律,保障 AI 技术与商业协同发展,重实干、重过程和重阶段,承担机会风险而不是技术瓶颈风险。技术瓶颈的突破可由研究机构和部门借助政府基金和高风险研发资本先行攻关实现。

第六,共享服务平台与大数据信息,实现城市发展的智慧化、特色化和均衡化。全国范围内,应该拆除各类公路收费站、取消各类通信区域限制(长途电话)、升级通信技术服务(5G 技术)、完善各级政府办公及政策信息和各类企业生产与服务信息。城市群范围内,在交通、住房和相关配套服务都已逐步完备的条件下,需要通过现代通信和大数据技术获取各类在线市民声音,改进各类民生服务,逐步实现城市群内的各类信息智慧化共享。在此基础上实现城市发展的特色化和均衡化:供应链上,大型城市发展企业集团总部,中型城市发展企业分部,小型城市发展零部件生产基地;产业升级上,相对发达的城市可以倾向于发展芯片、新材料、精密加工等高端研发和制造产业,相对落后的城市可以优先发展人工智能应用、大数据分析、软件外包等轻、快、高产业。

参考文献:

[1] 黄 虹,张恩焕,孙红梅,等. 融资融券会加大投资者情绪对股指波动的影响吗? [J]. 中国软科学,2016(3):151-161.

[2] Hong H, Xu D, Wang G A, et al. Understanding the determinants of online review helpfulness: A meta-analytic investigation [J]. Decision Support Systems, 2017, 102: 1-11.

[3] Pournarakis D E, Sotiropoulos D N, Giaglis G M. A computational model for mining consumer perceptions in social media [J]. Decision Support Systems, 2017, 93: 98-110.

[4] Komorowski M, Huu T D, Deligiannis N. Twitter data analysis for studying communities of practice in the media industry [J]. Telematics and Informatics, 2018, 35(1): 195-212.

[5] Ruan Y F, Durresi A, Alfantoukh L. Using Twitter trust network for stock market analysis [J]. Knowledge-Based Systems, 2018, 145: 207-218.

[6] Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market [J]. Journal of Computational Science, 2011, 2(1): 1-8.

[7] 戴德宝,薛 铭. 民营银行准入监管的演化博弈分析[J]. 上海经济研究,2017(9):34-46+58.

[8] 戴德宝,薛 铭. 信息不对称下民营银行市场准入监管的博弈研究[J]. 财会月刊,2017(15):114-118.

[9] 南晓莉. 新媒体时代网络投资者意见分歧对 IPO 溢价影响—基于股票论坛数据挖掘方法[J]. 中国软科学,2015(10):155-165.

[10] Thien Hai N, Shirai K, Velcin J. Sentiment analysis on social media for stock movement prediction [J]. Expert Systems with Applications, 2015, 42(24): 9603-9611.

[11] Fama E F. Market efficiency, long-term returns, and behavioral finance [J]. Journal of Financial Economics, 1998, 49(3): 283-306.

[12] Proellochs N, Feuerriegel S, Neumann D. Negation scope detection in sentiment analysis: Decision support for news-driven trading [J]. Decision Support Systems, 2016, 88: 67-75.

[13] Garcia-Medina A, Sandoval L, Banuelos E U, et al. Correlations and flow of information between the New York Times and stock markets [J]. Physica a-Statistical Mechanics and Its Applications, 2018, 502: 403-415.

[14] Oliveira N, Cortez P, Areal N. The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices [J]. Expert Systems with Applications, 2017, 73: 125-144.

[15] 许启发,伯仲璞,蒋翠侠. 基于分位数 Granger 因果的网络情绪与股市收益关系研究[J]. 管理科学,2017(3):147-160.

[16] Checkley M S, Higon D A, Alles H. The hasty wisdom of the mob: How market sentiment predicts stock market behavior [J]. Expert Systems with Applications, 2017, 77: 256-263.

[17] Ho C S, Damien P, Gu B, et al. The time-varying

nature of social media sentiments in modeling stock returns [J]. Decision Support Systems, 2017, 101: 69-81.

[18] 段江娇，刘红忠，曾剑平. 中国股票网络论坛的信息含量分析 [J]. 金融研究，2017(10): 178-192.

[19] 部　慧，解　峥，李佳鸿等. 基于股评的投资者情绪对股票市场的影响 [J]. 管理科学学报，2018(4): 86-101.

[20] 胡昌生，陶　铸. 个体投资者情绪、网络自媒体效应与股票收益 [J]. 预测，2017(3): 50-55.

[21] Da Z, Engelberg J, Gao P. Insearch of attention [J]. Journal of Finance, 2011, 66(5): 1461-1499.

[22] Yu Y, Duan W, Cao Q. The impact of social and conventional media on firm equity value: A sentiment analysis approach [J]. Decision Support Systems, 2013, 55(4): 919-926.

[23] Ranco G, BordinoI, Bormetti G, et al. Coupling news sentiment with web browsing data improves prediction of intra-day price dynamics [J]. Plos One, 2016, 11(1): e0146576.

[24] 黄润鹏，左文明，毕凌燕. 基于微博情绪信息的股票市场预测 [J]. 管理工程学报，2015, 29(1): 47-52.

[25] 张信东，原东良. 基于微博的投资者情绪对股票市场影响研究 [J]. 情报杂志，2017, 36(8): 81-87.

[26] 杭品厚. 软计算技术在金融预测应用进展研究 [J]. 金融理论与实践，2018(5): 101-108.

[27] 李宝仁，胡　蓓，陈相因. 投资者情绪与股票收益的实证分析——基于上证投资者情绪综合指数 [J]. 北京工商大学学报社会科学版，2012, 27(4): 91-97.

[28] 张成功，刘培玉，朱振方等. 一种基于极性词典的情感分析方法 [J]. 山东大学学报(理学版)，2012, 47(03): 47-50.

[29] Granger C W J. Investigatingcausal relations by econometric models and cross-spectral methods [J]. Econometrica, 1969, 37(3): 424-438.

(本文责编: 辛　城)